Longitudinal Analysis of Root Server Anycast Inefficiencies

Zhihao Li Dave Levin Neil Spring Bobby Bhattacharjee
University of Maryland

ABSTRACT

Anycast is widely used in critical Internet infrastructures, including root DNS servers, to improve their scalability, resilience, and geographic proximity to clients. In practice, anycast depends on interdomain routing to direct clients to their "closest" sites. As a result, anycast's performance is largely a result of available BGP routes.

We provide what we believe to be the first longitudinal study of how anycast performs in providing load balancing and geographic proximity. We examine about 400M queries per day collected from over 100 anycast sites of the D root DNS server for over a year. From this data, we find evidence of excessively unbalanced load: several anycast sites absorb the majority of the total traffic. Moreover, we find that most of the clients do not use geographically proximal sites; queries on average travel twice the minimum distance.

To investigate the root cause of these inefficiencies, we use more than 9,000 probes in RIPE Atlas to measure 9 out of 13 DNS root servers. We show two main causes for poorly balanced load and long query distance: insufficient peering between the hosting domain and large ISPs, and misconfigured routes, often due to route leakages, from ISPs with poor peerings.

1. INTRODUCTION

The DNS root servers are a uniquely critical part of the Internet infrastructure: they "bootstrap" the entire domain lookup hierarchy, serving as a point of last resort if a name server encounters a name about which it has no information. The DNS architecture accounts for their importance, and historically, root name service has been distributed over thirteen root name servers, each referred to by a "letter", A-root through M-root. Over the last two decades, most of these individual root servers have been anycasted, enabling replication of service across hundreds of replicas across the globe.

Caching in pervasive in DNS, and common names are usually cached at name servers. Other than zones that the root servers themselves serve (e.g. in-addr.arpa for reverse lookups), most queries at root servers are for uncommon domain names, or for names that are a result of software or user mistakes. Root servers can uniquely confirm that faulty top-level domains are indeed incorrect, and software is gated until such response is generated. Hence it is important that root servers can be reached with low latency. The privileged position of root servers in the name hierarchy also makes them an obvious target of attack, of which there are many on a daily basis. The thirteen lettered root servers, and their hundreds of replicas, are deployed with the goal of both increasing resilience and reducing access latencies.

In this paper, we revisit the question of how well does global anycast work in context of the root servers. Our results, derived from a longitudinal analysis of one year of daily traces from the D-root replicas and augmented with active measurements to the 9 roots show that query traffic is routinely directed to replicas thousands of miles away from the nearest available. Indeed, our results tend to reaffirm measurements from over a decade ago, which showed that F- and K-root servers incur an extra distance of 1000km over optimal for about 40% of the queries [22].

Why is it that after a decade of new anycast deployments of literally hundreds of new replicas, performance has not improved? Why are clients being systematically directed to replicas that are thousands of miles away, when there are root servers in the same city? Our goal is to provide a root-cause analysis of these anomalies, and identify specific interactions between the routing and DNS infrastructure that lead to sub-optimal performance, regardless of how many replicas are deployed.

In our analysis, we use geographic distance traversed by a query as the measure of goodness of the underlying anycast routing. This is in contrast to prior work, which compares the anycast latency compared to unicast to individual replicas. Unfortunately, comparing anycast latency to unicast couples their performance to underlying routes that are available. As an example, suppose that a root operator deploys ten new replicas distributed across the planet. But if the operator peers with a global ISP only in one location, then *all* of that ISP's traffic to that root server, no matter where it originates, must first traverse to the peering point, and then to one of the many replicas. We quantify this type of “funneling” effect that forces routes to root servers from major (Tier-1) ISPs to traverse pinch points in the topology, in many cases obviating the benefit of new replica deployments.

Structural deficiencies in the peering topology serve to explain long term inefficiencies. However, our data also shows sudden sharp spikes in distance traversed by queries, e.g., all queries from major European ISPs suddenly abandoning an Amsterdam replica for one in Chicago. The situation often remains “broken” for months on end, until it is fixed, just as abruptly. We trace such impulsive behavior to what almost certainly are misconfigurations or route leaks in BGP. We surmise that the amount of traffic misdirected by such misconfiguration is little enough, and that the DNS infrastructure is robust enough, that such anomalies can persist for months on end. Our longitudinal analysis enables us to trace the begin and end of multiple such events.

In investigating route inefficiencies, we focus on queries that originate in or traverse Tier-1 ISPs. Our results show that these queries often incur large distance penalties, often due to the “funneling” behavior described above. The underlying cause varies: at times it is due to the lack of sufficient number of peering points with the root server hosting domain; in other cases, traffic is funneled due to poor route configurations, and deployed replicas and peering points go unused. Thus, our fundamental mental model of anycast—providing a combination of reasonably good replica selection and reasonable load balance—fails to apply where it perhaps matters most. Our data show many examples where Tier-1 ISPs make the wrong choice about whether to carry root traffic to a distant customer ISP versus delivering it to a nearby replica.

Our contributions are the following:

- We provide a year long longitudinal analysis of traffic at the D-root server, quantifying the effectiveness of and change in anycast behavior as seen by 105 replicas. This part of our analysis shows the performance of anycast as experienced by DNS resolvers.

- Based on our data, and on measurements using RIPE Atlas probes to other root letters, we present a longitudinal analysis of the load (im-)balance across different replicas of multiple root servers. Our analysis tracks where queries originate (and thereby, where they *should* ideally go, and where they are directed to by the underlying anycast. This analysis provides a server-centric evaluation of the load balancing aspect of anycast.

- Finally, for many routes that are stable but inefficient, and also for events that cause routing to root replicas to change appreciably, we provide a BGP-based analysis of the underlying cause. Our analysis shows events that demonstrate the “route funneling” effect that directly leads to longer routes, and shows persistent misconfigurations in BGP. Conversely, some of the events we analyze rectify these problems leading to better performance across major ISPs. This part of our analysis correlates anycast inefficiencies to events at the network routing layer.

The rest of this paper is organized as follows: We review related work in Section 2, and present our datasets and measurement methodology in Section 3. In Section 4, we characterize the longitudinal performance of anycast over the span of a year; this analysis exposes pervasive inefficiencies in anycasted root servers. We analyze the root causes of these inefficiencies in Section 5. Finally, in Section 6, we conclude.

2. RELATED WORK

The work in this paper is generally related to systems that use anycast, and to the performance of root DNS servers. Much prior work [5, 25, 20] has evaluated the performance of root servers, without necessarily a focus on the deployed anycast infrastructure.

IP anycast [26, 17] is a widely used technique that allows services to be transparently replicated across the Internet. *Replicas* may be located at multiple *sites*; all replicas share the same public IP address, which is advertised using BGP from each site. Clients connect to the public IP address, and are “routed” to a replica based on the underlying BGP routes. While used in many applications, notably CDNs [13, 7, 11, 1, 6, 10], root DNS anycast is distinguished due to its intrinsic importance in the Internet protocol stack, and by the fact that multiple independent organizations use anycast to provide the overall service.

Our work builds off of over a decade of measurement studies on root DNS anycast performance. At a high level, our work differs from prior DNS performance evaluations in that we offer a longitudinal view of root DNS anycast performance, and we empirically evaluate root causes for performance inefficiencies.

RTT-based Performance of Root Anycast Several studies have compared the RTTs between clients and their anycast instances to the *smallest* RTT among all of the possible anycast instances [32, 9, 21, 8, 2].

In 2006, Sarat et al. [32] performed such RTT measurements using pings from PlanetLab hosts to analyze F- and K-roots. Anycast to K-root is also measured in [8] with an evaluation of the marginal benefit of individual replicas. In 2013, Liang et al. [21] applied the King RTT inference technique [14] to measure latencies between nearly 20K open resolvers and root and top-level DNS servers. Most recently, in 2016, Schmidt et al. [9] used RIPE atlas probes to measure RTTs to all

DNS root servers that support anycast. Across a decade of such measurements, there has been a relatively consistent finding: that, with respect to RTTs, the achieved (anycast) performance is close to the ideal (unicast) performance. Moreover, like with other anycast-related studies, these papers speculate that BGP routing has an impact on whether clients obtain their optimal instance.

Our work differs from these prior studies in two fundamental ways. First, we do not use RTTs as a metric for anycast performance, but rather use the geographic distance between clients and anycast instances. While this may seem like a subtle distinction, it is critical for exposing the impact that routing inefficiencies have on anycast: at a high level, if the minimum-RTT path is subject to the same routing issues as the anycast path, then comparing RTTs would mask the inefficiency. Indeed, our results paint a far less rosy picture of anycast performance than RTTs alone would seem to indicate. Second, we apply this insight into routing inefficiencies to empirically evaluate the impact that routing has on anycast.

Distance-based Analyses of Root Anycast Other studies have also used the relative geographic distance as a metric for comparing how well anycast chooses among instances. In 2006, Liu et al. [22] used two days' passive DNS data from C-, F-, K-root, and reported median additional distances of 6000 km, 2000 km, and 2000 km, respectively. For C-root, they found that over 60% of clients traveled an extra 5000 km longer than strictly necessary; for F- and K-roots, 40% of clients traveled an extra 5000 km. Kuipers [19] performed a somewhat coarser-grained analysis of 10 minutes of K-root's anycast performance, showing that most clients are not getting routed to their geographically closest anycast instance.

By comparison, we present a longitudinal study of anycast performance, spanning an entire year's worth of D-root query trace data, augmented with RIPE Atlas measurements to other root servers. We find that anycast can be highly dependent on a relatively small number of route changes, and thus that short-lived studies risk being non-representative. Interestingly, however, many of these broader trends—clients being sent to anycast instances thousands of kilometers farther away than their closest replicas—continue today, over a decade since some of these initial findings, despite the fact that hundreds of more replica sites are now available. This motivates our study into the root causes of these inefficiencies.

Route stability Due to its reliance on BGP, anycast measurements provide an implicit analysis of the stability of underlying BGP routes. This observation has been used in the context of root server anycast [3, 4] to understand the causes of route instability. Our focus in this work is different, in that we evaluate the goodness of the replica selected, regardless of how often clients change replicas. Our detailed analysis of specific events (Section 5.3) shed light on events that cause clients to be switched to geographically distant replicas. Our analysis focuses on cases when large numbers of queries are routed across oceans, when replicas are available within the same country or even the same city.

Root	**Operator**	**# of replicas**	**# sites (# global)**
A	Verisign Inc.	5	5 (5)
B	ISI	1	1 (unicast)
C	Cogent Comm.	8	8 (8)
D	Univ. of Maryland	114	108 (20)
E	NASA (Ames)	72	71 (13)
F	ISC Inc.	140	137 (5)
G	US Dept. of Defense	6	6 (6)
H	US Army	6	2 (2)
I	Netnod	59	53 (53)
J	Verisign Inc.	129	113 (67)
K	RIPE NCC	53	49 (48)
L	ICANN	160	142 (142)
M	WIDE Project	8	5 (4)

Table 1: Root server data, current as of May 2017. Data from `http://root-servers.org`.

Root Resilience The root servers are a seemingly favorite target for denial of service attacks [27, 31, 33, 15], and many papers [23, 34, 18] have investigated how anycasting provides resilience for the root DNS service. While we do not specifically evaluate attack resilience, our analysis shows that due to underlying routing imbalances, query load is highly non-uniform across replicas. Hence, a determined attack may be able to undermine the availability of root replicas by overwhelming the popular replicas. Due to the "funneling" effect of the underlying routes, deployed replicas may go unused even as the service as a whole is degraded due to attack.

As described in the next section, we base a part of our analysis on data from RIPE Atlas probes that utilize a particular type of DNS query which the identification of specific replicas. Prior research [16, 4, 19] had used the same technique to identify root replicas and understand client-replica mappings and measure service availability.

3. MEASUREMENT METHODOLOGY

Table 1 lists the number of locations, and the number of global and local sites for each root server in May 2017. In the rest of this section, we describe our measurement methodology and the resultant data sets.

3.1 Passive measurements

D-root is operated by the University of Maryland. As of May 2017, it has 108 anycast sites, 20 of which are global and the rest local. We capture 20% of all traffic at each replica, and base our analysis on this longitudinal data. For this paper, we consider the data collected on

Probe Location	# probes
Germany	1293
United States	1017
France	818
United Kingdom	621
Netherlands	541
Russia	481
Italy	307
Czechia	272
Switzerland	265
Canada	206

Table 2: Distribution of popular RIPE-Atlas probe locations.

every day in 2016. On average, in 2016, D-root received more than 30,000 queries per second resulting in about 140 GB of trace data per day.

3.2 Correlating Datasets

The passive D-root dataset gives us a global, detailed view of client activity seen at one root server. In order to correlate this view with other servers, and to understand underlying network dynamics, we augmented the D-root data results with active measurements that the RIPE Atlas probes conduct to root servers.

The RIPE Atlas framework [29, 30] is a set of ~9000 probes distributed across 179 countries and in ~3390 ASes as of May 2017. Each probe periodically executes pre-defined measurements, called "build-in measurements", that include specific DNS queries and traceroutes to all 13 DNS roots.

Our analysis includes measurements that the RIPE-Atlas probes sent to the 9 out of 13 roots that have at least 5 anycast global sites.[1] We focus on two specific measurements: DNS CHAOS queries and traceroutes.

DNS CHAOS queries retrieve data corresponding to the TXT record for the string "hostname.bind" with the DNS Class set to CHAOS (as opposed to Class Internet, which is the common case). The "hostname.bind" is a special record supported by BIND nameserver implementations, which is conventionally configured by the server operator to return a string that uniquely identifies the server replica. For instance, a representative DNS CHAOS response from D-root is "abva2.droot", which indicates that this is the second D-root replica located in Ashburn, Virginia. Similarly, C-root replicas return responses of the form "iad1a.c.root-servers.org", identifying replica "1a" located in Dulles, Virginia. Each RIPE-Atlas probe sends these queries to each root server's IP address (A-root through M-root) every 4 minutes. The underlying anycast mechanism (if the root server uses anycast) then directs the probe to a specific replica, which is then identified via the DNS CHAOS response. In our results, we use the data gathered by all probes over one year (Jan.–Dec., 2016); these measurements enable us to identify the specific replica that a particular RIPE-Atlas probe is directed to at a given time during 2016. We note that prior research [16, 4] had used the same technique to identify root replicas and understand client-replica mappings and measure service availability.

Along with DNS CHAOS queries, the RIPE-Atlas probes also conduct a traceroute to each root server's IP address every 30 minutes. The traceroute data enables us to map AS paths that were used during 2016 by the probes to reach different root servers.

Our D-root passive measurement data provides a global, unbiased sample of global DNS resolver distribution, both in terms of location and query volume. In contrast, the location of RIPE-Atlas probes are skewed towards locations in Europe, as shown in Table 2. Our analysis of D-root data (Table 3) will confirm this bias in RIPE-Atlas probe locations. Hence, we are careful not to draw "global" conclusions based on the RIPE-Atlas probes: instead, we use the RIPE probes to confirm inefficiencies seen in the global D-root data set, and to map underlying routes taken by different probes at specific points in time.

4. ANYCAST PERFORMANCE ANALYSIS

In this section, we characterize the performance of anycast service for root name servers in terms of two features: the distance traveled by queries compared to the distance to the nearest (usable) replica, and the load balance to replicas relative to the balance if queries went to the (geographically) nearest replica.

Although network latency measures and geographic distance are not interchangeable, this section focuses on how common it is for queries to cross continents and leave cities that have global replicas. We consider this behavior to be problematic, and in the following section will consider features of network routing policy that lead to these inefficiencies.

4.1 Can queries travel less far to reach D-root?

Figure 1 shows three average distances associated with queries to D-root through 2016. At the top is the average distance of actual queries received, using the distance between replica and query source. In the middle, is what the average distance could be if, as one might hope, queries were delivered to the closer of their current replica and any global replica. (This allows a client to continue using a local replica if that remains better than an alternative global replica.) At the bottom is the average distance if every query were permitted to go to the closest site, including "local" sites, representing an optimum given the query load and physical deployment, with unconstrained routing.

[1] We were unable to include measurements from G-root since it does not respond to "hostname.bind" DNS CHAOS queries with meaningful identifiers that we use to distinguish replicas.

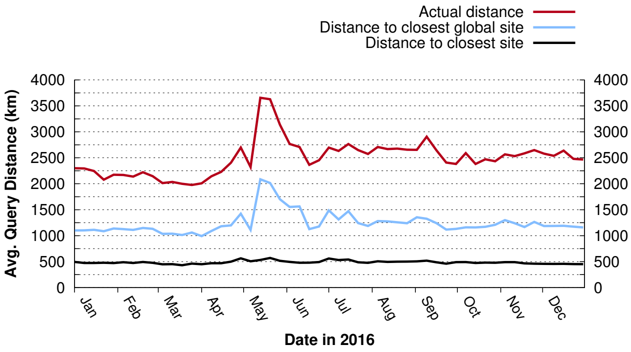


Figure 1: Average query distance to D-root in 2016.

Two key conclusions can be drawn from the changes in this graph: the top line increases, from 2250 km in January to 2500 km in December, showing that in practice, the average query is traveling farther, conversely, while the bottom line decreases slightly, reinforcing that this degradation in performance is despite adding replicas. Some dynamics to the graph are a feature of query load: the spike in mid-May is due to a heavy volume of new queries from South America that were routed to replicas in California and Florida. Although considering performance in terms of queries (rather than measurement sources) guides our effort to optimize, the volume (how many) and composition (where queries are from) is dynamic. This dynamism is facilitated in part by resolver implementations that will query a chosen letter persistently rather than round-robin among them: these resolver implementations should help to increase the number of short-distance queries as they make performance-based decisions, but may also cause discontinuities if they abandon D-root.

Note further that the average query can be answered with a round trip distance of 1000km: about 5 ms. On average, queries travel five times as far, leading to 25ms. From Figure 1, we conclude that there is substantial room for improvement in how anycast queries are directed to replicas.

4.2 Query distance distribution

While Figure 1 showed that average geographic distance for queries is substantially and persistently larger than necessary, the next step is to determine whether this average is drawn by a few pathological outliers that might be corrected or is widespread across small and large increases.

Figure 2 shows the distribution of query distances in the first week of October 2016. As in Figure 1, we consider both an unconstrained "closest site" optimal, a partially constrained "closest global" (or current local, if better), and the actual distance.

The median distance for a query is 1276 km. Yet, over 92% of queries have at least one replica within this distance, and 73% of queries have a global replica within this distance. Worse, about 5% of queries traveled farther than 10,000 km.

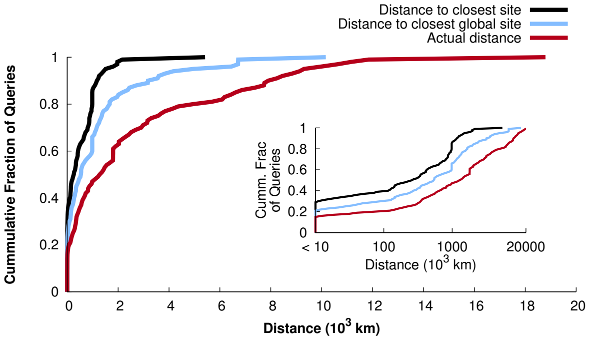


Figure 2: Distance traveled by queries received at D-root, October 1–7, 2016, compared to hypothetical best global and best overall replica distances. (The inset is the same plot with a log-scale x-axis.)

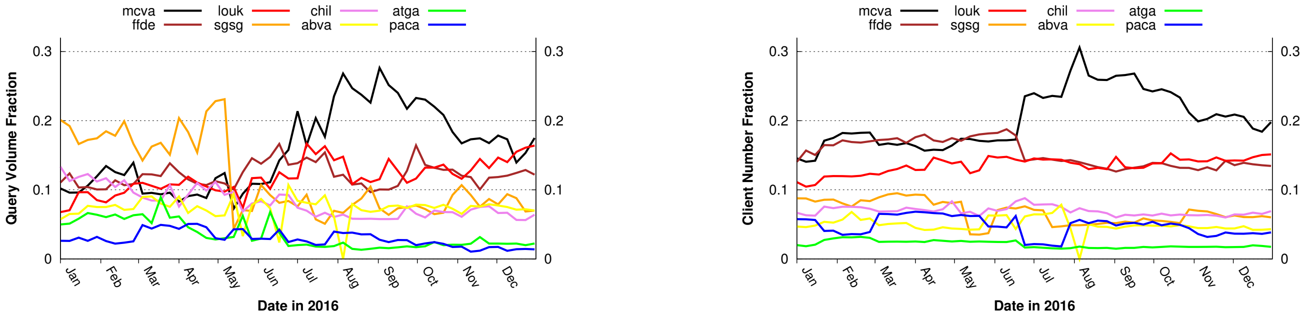
While Figure 2 shows that some queries travel very far, even queries that should be served by a nearby replica travel farther than necessary. As visible on the log scale in the inset of Figure 2, about 30% of queries are generated less than 10km away from a replica, yet only about 16% are served from within 10km.

4.3 Query and client load are out of balance

Next, we characterize the traffic load on D-root sites with two metrics: number of queries per day (Figure 3a) and the number of clients served per day (Figure 3b). We consider both metrics since, at times, the activity level of individual clients can vary tremendously: balancing to even out the number of clients served might induce an imbalance in queries. We do not assert that traffic should be perfectly balanced, and seek only to point out that it is well outside what one might expect in terms of both query load and client load. To analyze load imbalance requires a network trace dataset to understand the workload received by the root name server and cannot be completed with RIPE Atlas probes alone.

Overall, 90% of queries are answered by the 19 global sites, and the top 6 global sites serve 60% of the queries. (The top 6 are Frankfurt (*ffde*), London (*louk*), McLean Virginia (*mcva*), Singapore (*sgsg*), Ashburn Virginia (*asva*), and Chicago (*chil*).) "Local" sites account for the remaining 10%: they intentionally limit where queries may come from through *no-export* and *no-peer* attributes and selective peering. They are expected to have lower traffic load, and we focus only on the balance between global sites.

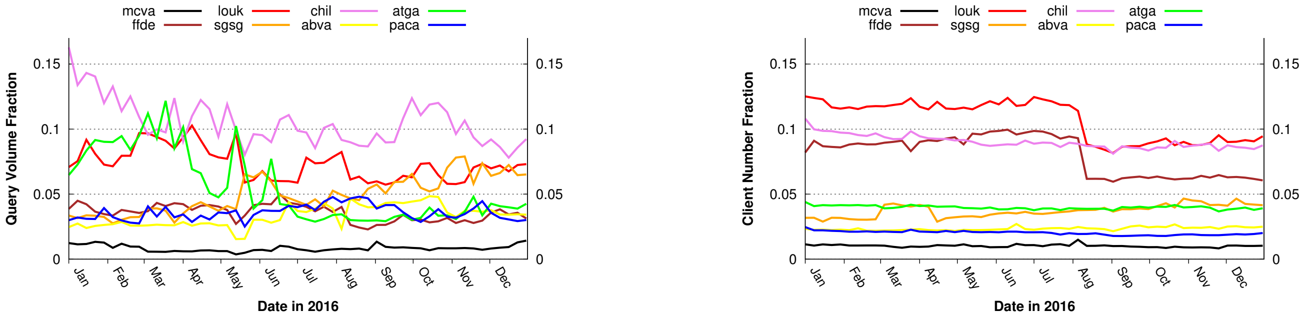
First, compare Figure 3a to Figure 3b. Interestingly, while query load to Singapore starts 2016 highest and drops in May, this change is not as dramatic in the number of clients. The top replica (*mcva*) increases its



(a) Query load among D-root sites.

(b) Clients served by D-root sites.

Figure 3: Fraction of queries received per day and number of clients served, for each week in 2016.



(a) Query distribution by closest global D-root site.

(b) Clients distribution by closest global D-root site.

Figure 4: Distribution by closest replica to query source for each query (left) and for each client (right), for each week in 2016.

load over the summer, peaking at 30% of the clients at one point in August, and finishing the year with 20%. For reference, if evenly balanced across 19 global sites, one would expect each metric to be just above 5%.

Next, contrast Figure 3a to Figure 4a. The latter shows the distribution of where queries should go, if directed to the closest global replica (including global replicas not shown). Notably, the most popular server (mcva) of Figure 3a, if choosing the nearest server, would be the least popular of this group. (This is partly due to having another very nearby replica in Ashburn, Virginia (*abva*). However, *abva* has a typical expected query load, reinforcing that *mcva* is an outlier.) Second, note that the query load to Singapore (*sgsg*) appears undeserved: that replica is not the closest for 20% of the queries.

Finally, we note two events that significantly reduce query distances in 2016. The drop in August in Figure 4b occurs when a new global replica is activated in Amsterdam, reducing the fraction of clients closest to London or Frankfurt. (The Amsterdam replica is not one of the top-8 replicas, and thus is not represented in the figure.) This did not seem to have a substantial effect on the fraction of queries that went to London or Frankfurt. However, the Chicago and London replicas both serve about twice as many clients as expected given 19 global replicas overall. Figure 3a shows that the query volume for the *sgsg* replica in Singapore drops sharply in May 2016. This is because many queries from China (sourced from China Telecom and China Unicom), which used to be routed to a global replica Singapore, are routed instead to a new *local* replica in Sydney, Australia. We speculate that this behavior is due to a BGP route leak from a ISP in Australia that peers with the Chinese ISPs. We are unable to confirm this however; in Section 5.3, we investigate multiple similar instances involving other large ISPs that we are able to confirm using BGP announcements.

Country	**Frac. of Queries**	**Avg. query dist. (km)**	**Avg. opt. dist. (km)**
United States	32.0%	1413.6	367.9
China	9.7%	5746.5	1413.8
Brazil	4.6%	7740.0	6483.3
Netherlands	3.6%	792.2	293.1
United Kingdom	3.4%	495.1	23.9
Germany	3.1%	938.0	128.3
Russia	2.7%	2984.4	1939.5
Japan	2.6%	1262.0	40.7
Canada	2.5%	943.1	429.5
Chile	2.1%	8738.9	6515.9

Table 3: Top ten locations with most queries per day sent to D-Root.

Provider	Code	ASNs
AT&T	AT&T	7018
Cogent Communications	COGENT	174
Deutsche Telekom AG	DTAG	3320
Global Telecom & Technology	GTT	3257, 4436
KPN	KPN	286
Level 3 Communications	LEVEL3	3356, 3549
Liberty Global	LGI	6830
MCI Communications	UUNET	701, 702, 703
NTT Communications	NTT	2914
Orange S.A.	OPENTRANSIT	5511
Qwest Communications	QWEST	209
Sprint	SPRINTLINK	1239
TATA Communications	TATA	6453
Telecom Italia	SEABONE	6762
Telefonica Network	TELEFONICA	12956
Telia Carrier	TELIANET	1299
XO Communications	XO	2828
Zayo Group	ZAYO	6461

Table 4: List of Tier-1 ISPs. The "Code" column lists the string by which the ISP is identified in our results.

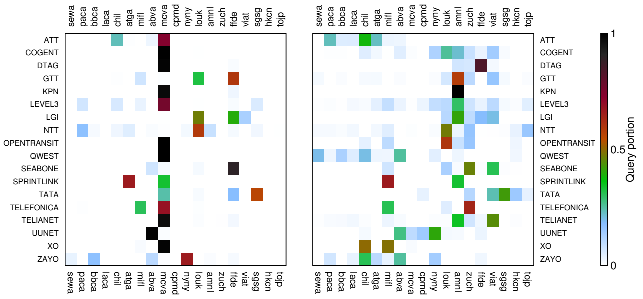


Figure 5: Distribution of queries routed by Tier-1 ISPs for D-root. The left panels shows which sites the queries went to; the right panel shows which sites are nearest the RIPE Atlas probes.

5. ROOT CAUSES OF INEFFICIENCY

Our longitudinal analysis of D-root shows that queries often traveled thousands of kilometers further than necessary, and that such problems were pervasive and persistent. These inefficiencies manifest themselves as load imbalances at root replicas, and are often a result of pathological conditions in the underlying BGP routes. We begin with an analysis of query load distribution for D-root replicas next. Section 5.2 shows similar analysis for other root servers, followed by analysis of how the measured performance exposes routing dynamics (Section 5.3).

5.1 Inefficiencies and Imbalances at D-root

In Section 4, we found that queries are routed far away from nearby global replicas. Hence, the paths chosen by the ISPs that are carrying this traffic must, at some level, be suboptimal. To investigate this hypothesis, we focus on how queries that are routed through Tier-1 ISPs (identified from the data in [12]) listed in Table 4) reach D-root replicas. Specifically, for each RIPE-Atlas probe that originates in or traverses a Tier-1 ISP, we catalog (1) the replica the query is directed to, and (2) the closest global replica the query *could* have been directed to, had the underlying BGP path existed and been chosen. (This analysis is possible since RIPE provides accurate probe locations, traceroutes of the paths they take to the chosen root replica, and identification of which replica is chosen at the time, enabling us to determine the ASs traversed by the queries.)

Prior to analyzing our results, we note two restrictions. First, we consider only global replicas, since local replicas are advertised within a single AS only, accounting for less than 10% of total D-root queries. Local replicas appear rarely for traceroute paths that traverse a Tier-1. Second, we only analyze paths through Tier-1 ISPs. This restriction allows us to understand how large ISPs, with networks that have a global footprint and many hundreds of peerings, interact with the DNS anycast. Tier-1 ISPs source over 5% of the queries in our passive D-root measurements, and roughly 35% of the RIPE-Atlas probes traverse a Tier-1 ISP. (We suspect that the fraction of queries that traverse a Tier-1 ISP before reaching D-root is about as high (35%) because a comparable fraction of RIPE probes originate in Tier-1 ISP address space (8%).) Thus, we believe restricting our analysis to Tier-1 ISPs provides a representative picture of global DNS anycast.

In our first set of results, we consider query routing using a snapshot of data collected by the RIPE-Atlas probes on October 1, 2016. Figure 5 contains two heatmaps. At left is a heatmap of global replicas to which queries from RIPE Atlas probes to D-root *were* directed on that day, arranged by Tier-1 ISPs traversed. (If a query traversed more than one Tier-1 ISP, the path is classified by the *first* ISP.) Darker shades represent higher query volume and the figure shows that most Tier-1 ISPs sent a large fraction of their traffic to the *mcva* or *abva* replicas. Although there are 20 global D-root replicas (the 20th was added in August 2016), the dark vertical line in this figure shows that most traffic is concentrated predominantly on one replica. Conversely, many replicas go virtually unused.

The right side of Figure 5 shows how the queries would be distributed if each query had been directed to its closest replica. The distribution at right is a rough approximation of the locations of RIPE Atlas probes hosted by networks that need a Tier-1 hop to reach D-root. This figure represents what IP anycast could *ideally* achieve, and it reflects what anycast's properties *ought* to be: a far more even distribution of load and more low-distance queries than what actually occurs. Indeed, it is the mismatch between the left and right side of Figure 5 (and in turn, the misdirection of queries) that causes the longer distances to traversed as discussed in Section 4.

These figures show many examples of pathological path length inflation: Deutsche Telekom, KPN, and Telianet direct most of their queries that originate in Europe to the *mcva* replica in Virginia, bypassing multiple European D-root sites in Frankfurt, Amsterdam, and London. Similarly, queries routed through Cogent, QWEST, Opentransit, UUNet and XO could benefit from being routed to closer sites, but generally get routed to *mcva*.

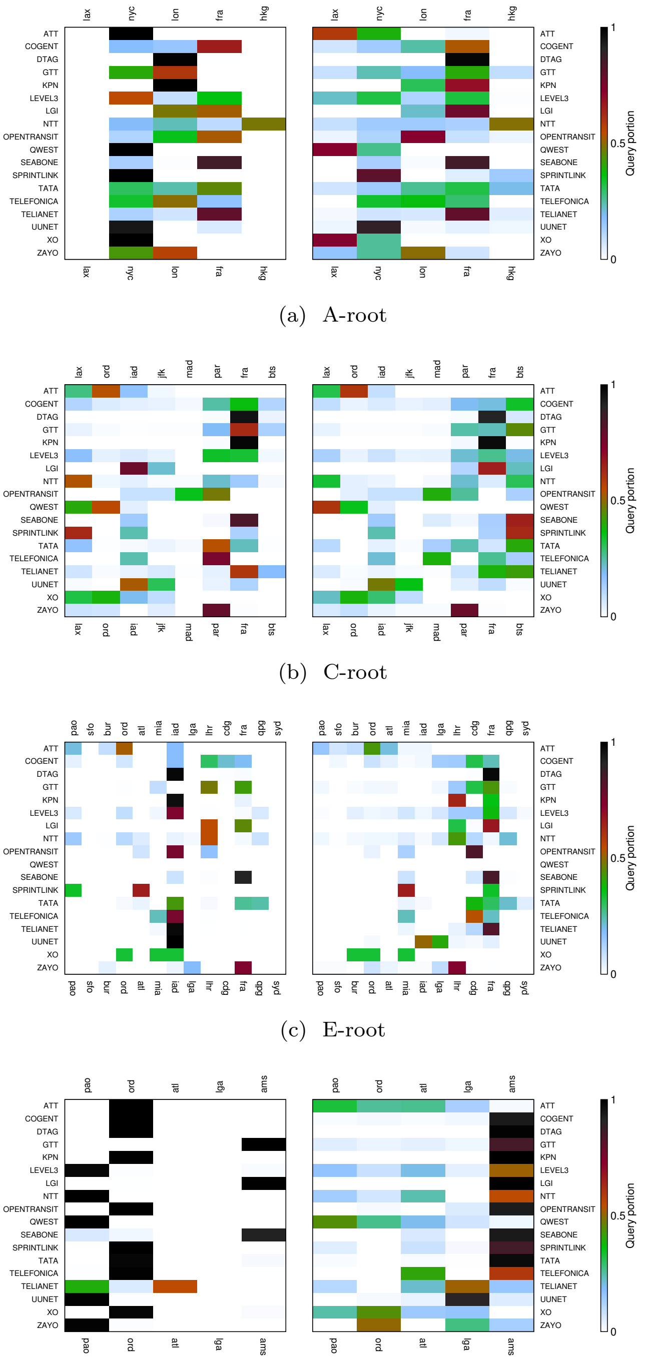
The explanation for *mcva*'s popularity lies in how the prefix is announced to BGP. The D-root prefix is broadly announced by Packet Clearing House (PCH, AS42), which routes queries to all replicas except *mcva*. The D-root prefix is also announced by MAX-GIGAPOP (AS10886), which directs all its queries to the *mcva* replica. Many Tier-1 ISPs (e.g., Cogent (AS174), LEVEL3 (AS3356) and QWEST (AS209)) directly peer with AS 10886, and hence queries that originate all across the globe are "funneled" to the Tier-1–MAX-GIGAPOP peering site, and eventually to *mcva*.

5.2 Other Roots

Using the same methodology (traceroutes on October 1, 2016, from RIPE-Atlas probes), we analyzed the same data for the eight other root servers that have at least five global anycast sites. Figure 6 repeats the style of Figure 5: pairs of "where queries went" versus "where the nearest replica is" for each Tier-1 ISP. These figures show that the poor behavior seen at D-root is not an exception, but the norm.

For instance, two out of five sites for A-root remain nearly unused by RIPE-Atlas probes. For A-root, Deutsche Telekom routes queries to London but not Frankfurt. E-root's query distribution is similar to D-root: a site in Virginia(*iad*) is preferred by many Tier-1 ISPs, including for queries that originate in Europe, bypassing replicas in London, Paris and Frankfurt. For F-root, the Chicago replica is preferred for many queries that should have been directed to Amsterdam. We performed the same analysis for I-, J-, K-, and L-root as well. All of them show similar preference for individual replicas, perhaps more notable only because each comprises dozens of global replicas.

C-root, however, is a notable exception. As Figure 6(b) shows, C-root matches queries to replicas far more effectively than other roots. Schmidt [9] observed a similar result: the median anycast latency for C-root over the minimum unicast latency was only 5 ms. Our result shows that C-root's good performance is not simply due to the paucity of unicast routes: C-root's operators peer with Tier-1s at sufficient numbers of peering points to provide the benefit of geographically replicated anycast. Perhaps unsurprisingly, C-root is itself operated by a Tier-1 ISP (Cogent), which peers with other Tier-1 ISPs widely. Queries to C-root that traverse other Tier-1s are "early-exit"-ed to Cogent, and then efficiently routed within the Cogent Tier-1 network.



(a) A-root

(b) C-root

(c) E-root

(d) F-root

Figure 6: Distribution of queries routed by Tier-1 ISPs for different root servers. The left panel shows which sites the queries went to (typically mcva); the right panel shows which sites are nearest the RIPE Atlas probes.

Our analysis shows mismatches between "optimal" routing and realized routing for almost all root servers, except for C-root, as noted above. These mismatches often lead to the query "funneling" effect, which manifests itself as a dark vertical line in the heatmaps. In these cases, most queries from the affected Tier-1 ISPs are routed to one or two replicas, regardless of where the queries originate. There are two primary reasons for such poor behavior: the first, as has been noted in prior work [9, 19], is that the anycasted domain does not peer sufficiently with providers, causing all queries to be routed through a few peering points. D-root shows an example of this behavior whereby AS10886 advertises the prefix for D-root, but peers with many Tier-1 providers only near the *mcva* replica, causing traffic from many Tier-1 ASs to be funneled to *mcva*. A somewhat more subtle cause of query funneling occurs regardless of how many peering points the hosting domain has. In fact, as the hosting domain peers with *more* providers, poor routing decisions upstream can lead to query funneling! We examine instances of such anomalous behavior next.

5.3 Performance changes expose routing dynamics

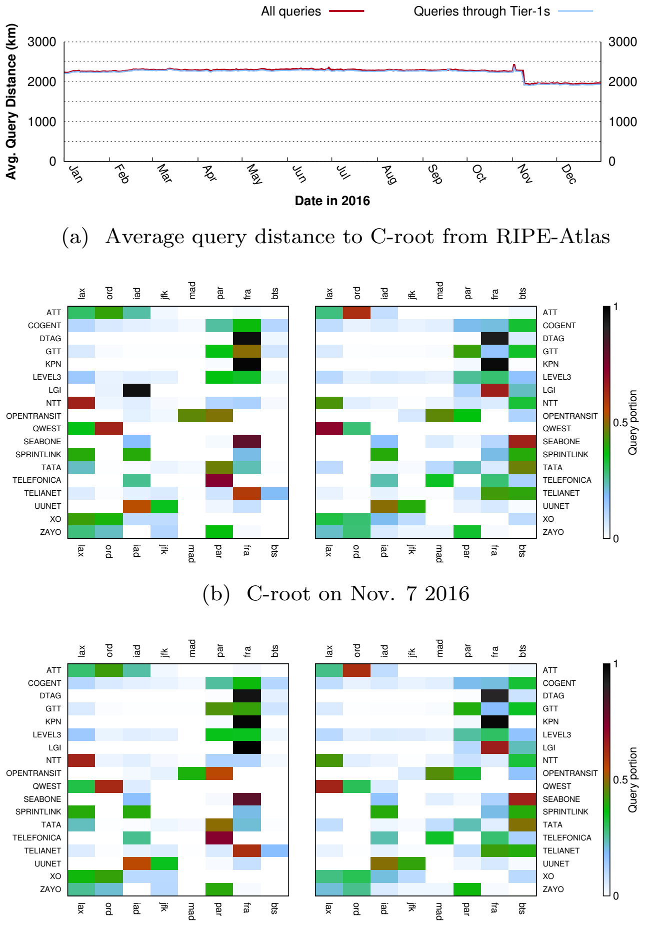
During 2016, the distance between RIPE Atlas probes and their chosen replicas was not consistent, showing substantial events. In this section, we consider changes that affected C, D, E, and F-root addresses at some point in the year. Events for D, E, and F, show clear changes in the provider-level AS path: a set of Tier-1 ISPs changed how they reached the root address, typically choosing a single poor replica. The C-root event does not manifest as a change in AS path, but BGP advertisement traffic supports that a significant routing change was made. We discuss these changes in alphabetical order.

At a high level, these results show that average query distances remain relatively stable for months, but show sudden impulsive behavior that can affect query distances by thousands of kilometers.

5.3.1 C-root: LGI chooses a better peering

In November 2016, Figure 7a shows that the average distance for queries to C-root decreased from 2300km to 2000km. (Because C-root is operated by Cogent, all queries traverse a Tier-1, meaning that the lines of the figure are overlapped; we use the difference to show the impact of routing changes beyond the Tier-1 ISPs.)

We next compare how Tier-1s routed queries the day before (Figure 7b) the change and the day after (Figure 7c). As in Figures 6 and 5, the left shows where queries went and the right shows which replica is nearest. The key difference between Figure 7b and 7c is that traffic from LGI is routed instead to Frankfurt (*fra*),



(a) Average query distance to C-root from RIPE-Atlas

(b) C-root on Nov. 7 2016

(c) C-root on Nov. 9 2016

Figure 7: Distribution of queries to C-root routed by Tier-1 ISPs before and after the routing change. (a) Average query distance over time, (b) Query distribution by first Tier-1 ISP before and (c) after, (d) AS paths evident in traceroutes before and (e) after. In the AS graphs, edges represent appearance in traceroutes from at least 4 sources, solid edges at least 15, and thicker edges at least 100.

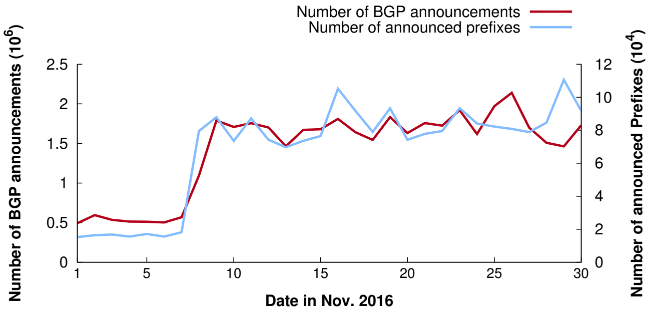


Figure 8: Number of BGP announcements through LGI-Cogent peering, early November 2016.

nearer to the clients that it supports.

Because C-root is operated by Cogent, and LGI peers directly with Cogent, we sought to confirm that there was a significant routing change that occurred. In IP address space, the paths clearly traverse a different set of IP addresses to cross the peering. In BGP, the volume of BGP traffic associated with LGI to Cogent increased significantly at the same time, as shown in Figure 8. This analysis uses BGPStream [24] to see BGP updates collected from RouteViews, focusing on prefixes advertised with the tuple LGI-Cogent (AS6830-AS174) in the AS Path. The plot shows that the number of announcements and prefixes with LGI-Cogent tripled around November 9, suggesting increased connectivity between the two.

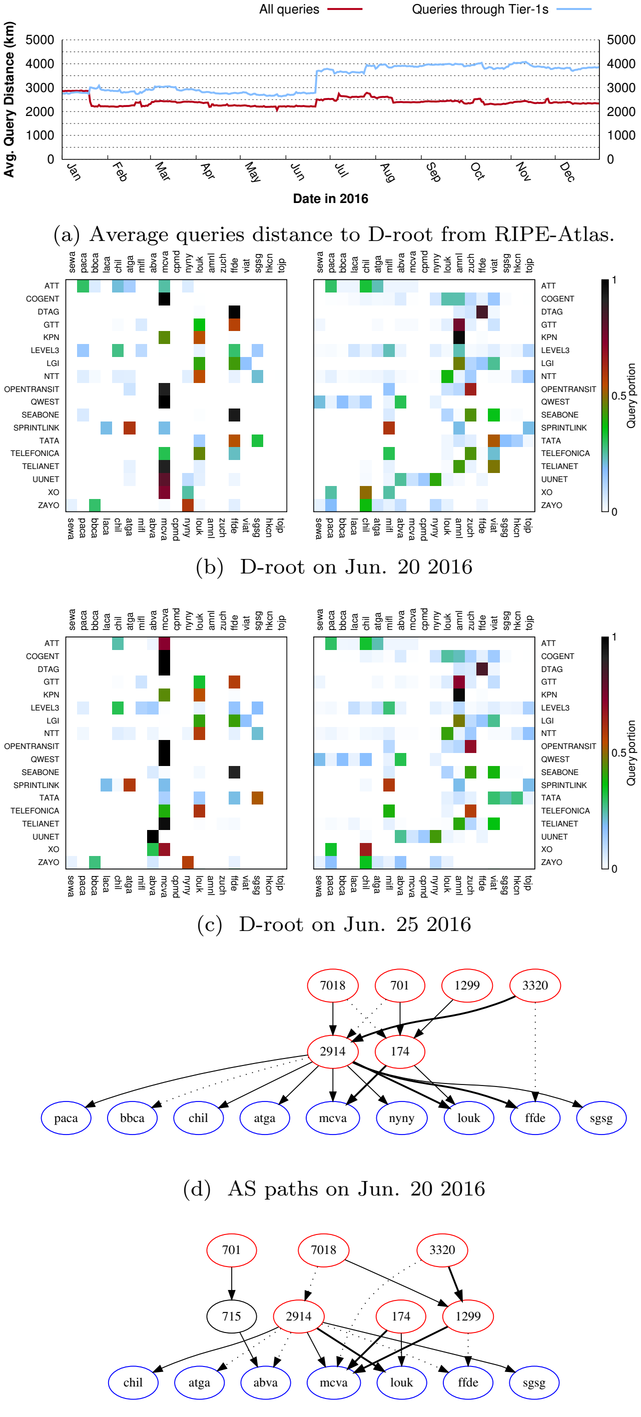
5.3.2 *D-root: Telia pulls DTAG to mcva*

In June 2016, Figure 9a shows that the average distance for queries to D-root increased by about 300km, or by about 1000km if considering only queries that traversed Tier-1 ISPs. The key difference between Figure 9b and 9c is a shift toward the *mcva* replica for DTAG.

Figures 9d, 10d, and 11d, described in more detail below, share a common dataset and format. The underlying dataset comprises traceroutes taken from RIPE Atlas probes to the root server's anycast address. Concurrently, RIPE probes query a special record from the root name server to determine which one was in use at the time. We translate the IP addresses of hops along the path into their originating AS to construct the traceroute-based AS path, then show only edges after the ISPs involved in route changes.

In the figures, numbered nodes indicate Tier-1 ASNs that were part of a routing change, or non-Tier-1 ASNs they used to reach a replica. Named nodes at the bottom indicate the replicas that were reached by this set of ISPs. In this set of changes, the number of replicas in use is reduced through the change. Line style and thickness indicates the number of traceroutes that included a link from one AS to another. No edge appears if fewer than four traceroutes included such a link. Edges appear dotted unless seen at least 15 times: on one hand, relatively few observations may be due to transient behavior, on the other, omitting these edges may hide diversity that does not happen to be observed by RIPE probes. Plain edges are up to 100 observations, then lines are slightly thicker to 800, and in one case where roughly 1/10 of all RIPE probes used the connection from 3356 to 3549, a thickest line.

Figure 9d and 9e shows before and after Telia (1299) provided a direct route to the *mcva* (northern Virginia) replica, rather than use Cogent (174). DTAG (3320) and AT&T (7018) switched routes to D-root from NTT (2914) to Telia (1299). Telia appears to direct most all



(a) Average queries distance to D-root from RIPE-Atlas.

(b) D-root on Jun. 20 2016

(c) D-root on Jun. 25 2016

(d) AS paths on Jun. 20 2016

(e) AS paths on Jun. 25 2016

Figure 9: Distribution of queries to D-root routed by Tier-1 ISPs before and after the routing change. (a) Average query distance over time, (b) Query distribution by first Tier-1 ISP before and (c) after, (d) AS paths evident in traceroutes before and (e) after. In the AS graphs, edges represent appearance in traceroutes from at least 4 sources, solid edges at least 15, and thicker edges at least 100.

traffic to the Northern Virginia (mcva) replica, and did so even before the event when it first traversed Cogent (174) address space.

The precise scenario is unclear, but this event would reinforce that, to avoid sending its own traffic far, a Tier-1 should not peer with an anycast operator in just one location. In the event that a single peering is desired, to avoid collecting traffic to be sent far, a Tier-1 should avoid exporting a route to others when having a connection to only one replica.

5.3.3 *E-Root: 3356 starts advertising a route*

In July 2016, Level3 appears to have begun treating an AS it acquired (3549) as a sibling, re-advertising the route to E-root, instead of as a peer where it would not re-advertise. This general change in relationship between 3356 and 3549 has been documented by Dyn research [28]. The impact of this change appears in Figure 10a, increasing the distance from RIPE Atlas probes to E-root by 800km, and for the subset of queries that traversed a Tier-1, 1500km.

Figure 10c shows that various providers switched from a replica that was appropriate for the client set (typically Frankfurt/*fra*) to northern Virginia (*iad*). Figure 10e shows the change to the AS path involved. Various providers that previously used NTT (2914) to reach E-root chose the new Level3 route, although 3356 directed those queries to a specific address within 3549, which then sent those queries to Northern Virginia (*iad*).

5.3.4 *F-Root: Comcast advertises a route to Chicago*

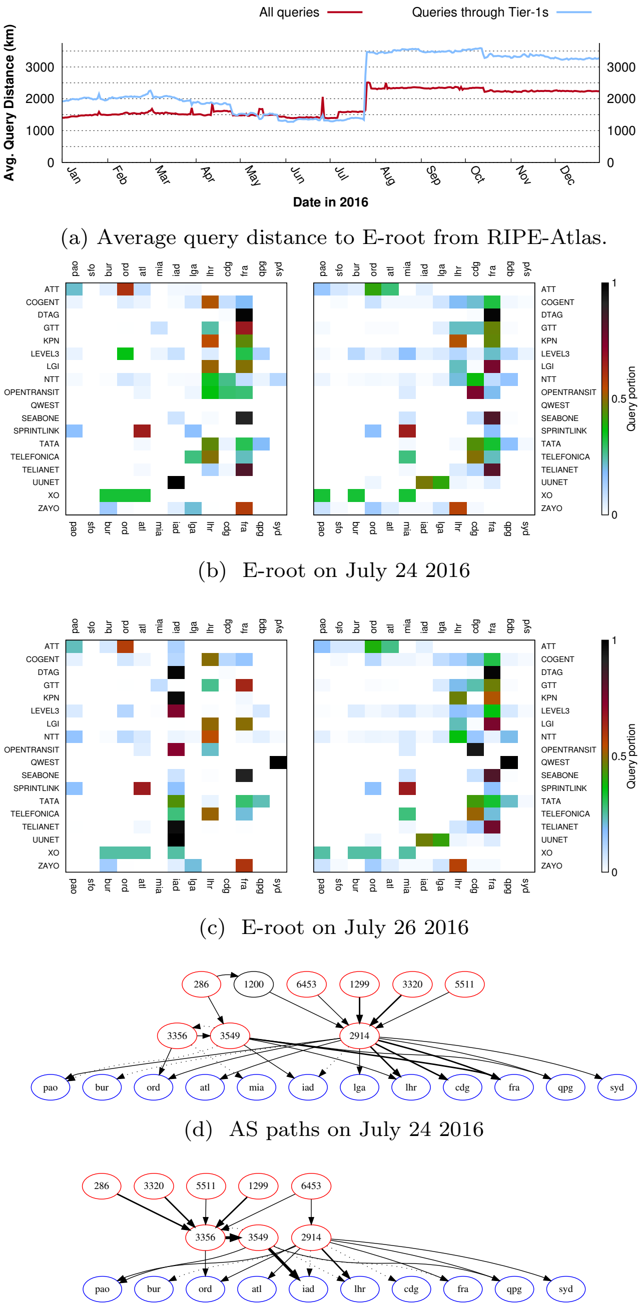
In March, the average distance to F-root increased by almost 1,300km, as shown in Figure 11a. This is the result of shifting substantial traffic to the Chicago replica, shown in Figures 10b and 10c.

Figures 11d and 11e show before and after Comcast (7922) appears to have advertised a route to F-root, despite delivering queries it received only to the Chicago (ORD) replica. Notable is the prior diversity of replicas (5 vs., in practice 1) and paths for this set of ISPs. 7922 may be seen as a customer by other ISPs, which could explain why so many Tier-1 ISPs chose the route to F-root through 7922.

In this plot, the middle tier (7922, 2914, 1280, etc.) are only shown for traceroute paths that traverse the ISPs above. For example, the connection from UUNET (701) to Palo Alto (PAO) appears over 100 times overall in the data, but appears only rarely in a 12956-to-701-to-pao path. This change was corrected in November 2016, as can be seen in Figure 11a.

6. CONCLUSION

We studied the anycast performance of DNS replicas through 2016, considering performance in terms of geographic distance relative to the nearest (global) replica,



(a) Average query distance to E-root from RIPE-Atlas.

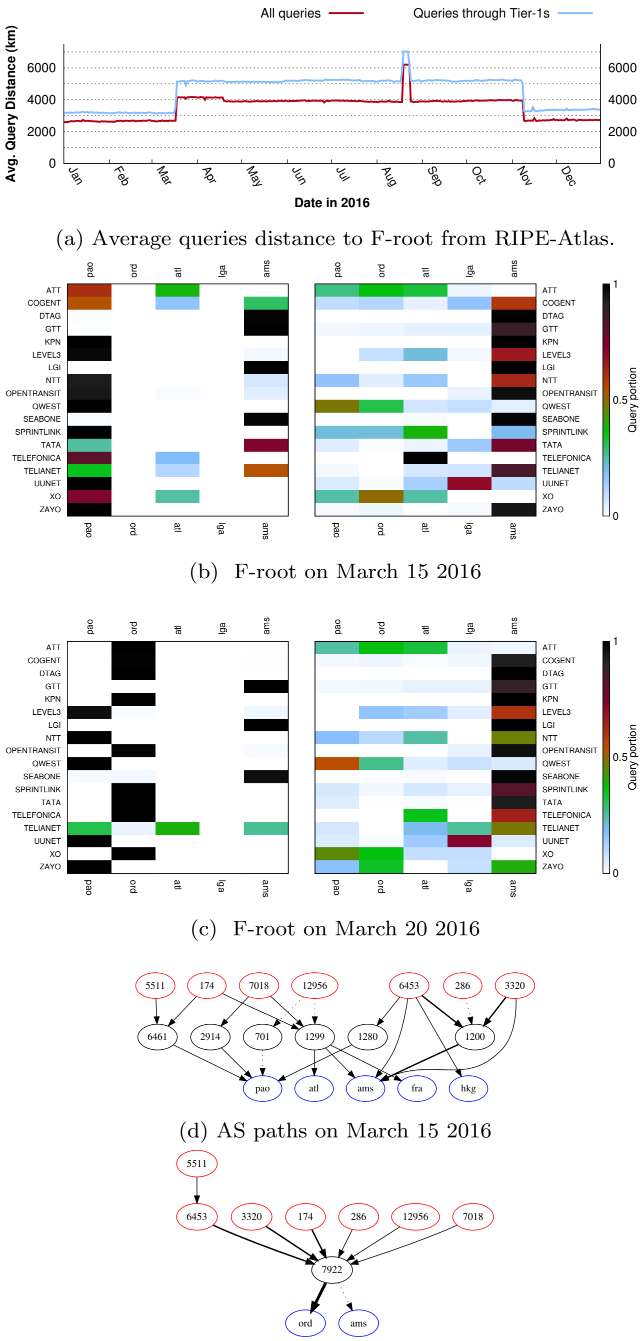
(b) E-root on July 24 2016

(c) E-root on July 26 2016

(d) AS paths on July 24 2016

(e) AS paths on July 26 2016

Figure 10: Distribution of queries to E-root routed by Tier-1 ISPs before and after the routing change. (a) Average query distance over time, (b) Query distribution by first Tier-1 ISP before and (c) after, (d) AS paths evident in traceroutes before and (e) after. In the AS graphs, edges represent appearance in traceroutes from at least 4 sources, solid edges at least 15, and thicker edges at least 100. Extra thickness represents the over 800 traceroutes that traversed the 3356 to 3549 and 3549 to iad replica links.



(a) Average queries distance to F-root from RIPE-Atlas.

(b) F-root on March 15 2016

(c) F-root on March 20 2016

(d) AS paths on March 15 2016

(e) AS paths on March 20 2016

Figure 11: Distribution of queries to F-root routed by Tier-1 ISPs before and after the routing change. (a) Average query distance over time, (b) Query distribution by first Tier-1 ISP before and (c) after, (d) AS paths evident in traceroutes before and (e) after. In the AS graphs, edges represent appearance in traceroutes from at least 4 sources, solid edges at least 15, and thicker edges at least 100. Extra thickness represents the over 800 traceroutes that traversed the 7922 to ord replica link.

and found substantial inefficiencies in how anycast queries are routed by many Tier-1 ISPs. This inefficiency is not an artifact of biased vantage point selection in RIPE Atlas probes, but appears as well in the passive traces of queries reaching D-root, persistently over the entire year. We then looked at the load balance across anycast replicas, and found that, again, the expected advantage of anycast is unrealized: queries are not well distributed across replicas. Finally, we described the routing path changes that affected the performance of accesses to C, D, E, and F root. The same behaviors affect the remaining anycast replicas in equal measure.

The C-root anycast infrastructure is largely resilient to routing problems, since various Tier-1 ISPs are encouraged to drop traffic to its operator, Cogent, using early exit routing. Yet, it is still possible to improve, as shown by the change in performance of queries from LGI. The performance of other root letters is vulnerable to misconfigurations in which routes are exported that are attractive for route selection (e.g., as a customer route) but ineffective for anycast distribution (e.g., concentrating on a single replica).

In future work, we intend to automate our analysis and identification of key routes that contribute to inefficient distribution, as well as to identify constructive changes that might improve overall performance through selective additional peering or corrections to how anycast routes are advertised and propagated.

7. REFERENCES

[1] H. A. Alzoubi, S. Lee, M. Rabinovich, O. Spatscheck, and J. Van Der Merwe. A practical architecture for an anycast CDN. *ACM Transactions on the Web (TWEB)*, 5(4):17, 2011.

[2] H. Ballani, P. Francis, and S. Ratnasamy. A measurement-based deployment proposal for IP anycast. In *Proceedings of the 6th ACM SIGCOMM conference on Internet measurement*, pages 231–244. ACM, 2006.

[3] B. Barber, M. Larson, and M. Kosters. Traffic source analysis of the J root anycast instances (talk). `https://www.nanog.org/meetings/nanog39/presentations/larson.pdf`, 2006.

[4] P. Boothe and R. Bush. DNS anycast stability. *19th APNIC,05*, 2005.

[5] N. Brownlee, K. Claffy, and E. Nemeth. DNS Root/gTLD performance measurements. *USENIX LISA, San Diego, CA*, 2001.

[6] M. Calder, X. Fan, Z. Hu, E. Katz-Bassett, J. Heidemann, and R. Govindan. Mapping the expansion of Google's serving infrastructure. In *ACM Internet Measurement Conference (IMC)*, pages 313–326. ACM, 2013.

[7] M. Calder, A. Flavel, E. Katz-Bassett, R. Mahajan, and J. Padhye. Analyzing the

performance of an Anycast CDN. In *ACM Internet Measurement Conference (IMC)*, pages 531–537. ACM, 2015.

[8] L. Colitti, E. Romijn, H. Uijterwaal, and A. Robachevsky. Evaluating the effects of anycast on DNS Root name servers. *RIPE document RIPE-393*, 6, 2006.

[9] R. de Oliveira Schmidt, J. Heidemann, and J. H. Kuipers. Anycast latency: How many sites are enough? In *Passive and Active Network Measurement Workshop (PAM)*, pages 188–200. Springer, 2017.

[10] X. Fan, E. Katz-Bassett, and J. Heidemann. Assessing affinity between users and CDN sites. In *International Workshop on Traffic Monitoring and Analysis*, pages 95–110. Springer, 2015.

[11] A. Flavel, P. Mani, D. A. Maltz, N. Holt, J. Liu, Y. Chen, and O. Surmachev. Fastroute: A scalable load-aware anycast routing architecture for modern CDNs. *connections*, 27:19, 2015.

[12] C. for Applied Internet Data Analysis (CAIDA). AS Relationships Dataset. `http://www.caida.org/data/as-relationships/`.

[13] D. Giordano, D. Cicalese, A. Finamore, M. Mellia, M. MunafÃ, D. Z. Joumblatt, and D. Rossi. A first characterization of anycast traffic from passive traces. IFIP, 2016.

[14] K. P. Gummadi, S. Saroiu, and S. D. Gribble. King: Estimating latency between arbitrary Internet end hosts. In *ACM Internet Measurement Workshop (IMW)*, 2002.

[15] ICANN. Root server attack on 6 February 2007. `https://www.icann.org/en/system/files/files/factsheet-dns-attack-08mar07-en.pdf`, 2007.

[16] D. Karrenberg. Anycast and BGP stability: a closer look at DNSMon (talk). `http://meetings.ripe.net/ripe-50/presentations/ripe50-plenary-tue-anycast.pdf`, 2005.

[17] D. Katabi and J. Wroclawski. A framework for scalable global IP-anycast (GIA). *ACM SIGCOMM Computer Communication Review*, 30(4):3–15, 2000.

[18] J. Kristoff. Feb 6/7 2007 DNS attack recap (talk). `https://www.dns-oarc.net/files/dnsops-2007/Kristoff-Feb07-attacks.pdf`, 2007.

[19] J. H. Kuipers. Analyzing the K-root DNS anycast infrastructure. 2015.

[20] B.-S. Lee, Y. S. Tan, Y. Sekiya, A. Narishige, and S. Date. Availability and effectiveness of Root DNS servers: A long term study. In *Network Operations and Management Symposium (NOMS), 2010 IEEE*, pages 862–865. IEEE, 2010.

[21] J. Liang, J. Jiang, H. Duan, K. Li, and J. Wu. Measuring query latency of top level DNS servers. In *Passive and Active Network Measurement Workshop (PAM)*, pages 145–154. Springer, 2013.

[22] Z. Liu, B. Huffaker, M. Fomenkov, N. Brownlee, et al. Two days in the life of the DNS anycast root servers. In *Passive and Active Network Measurement Workshop (PAM)*.

[23] G. Moura, R. d. O. Schmidt, J. Heidemann, W. B. de Vries, M. Muller, L. Wei, and C. Hesselman. Anycast vs. DDoS: Evaluating the November 2015 root DNS event. In *Proceedings of the 2016 ACM on Internet Measurement Conference*, pages 255–270. ACM, 2016.

[24] C. Orsini, A. King, D. Giordano, V. Giotsas, and A. Dainotti. BGPStream: a software framework for live and historical BGP data analysis. In *ACM Internet Measurement Conference (IMC)*, pages 429–444. ACM, 2016.

[25] J. Pang, J. Hendricks, A. Akella, R. De Prisco, B. Maggs, and S. Seshan. Availability, usage, and deployment characteristics of the Domain Name System. In *ACM Internet Measurement Conference (IMC)*, pages 1–14. ACM, 2004.

[26] C. Partridge, T. Mendez, and W. Milliken. Host anycasting service. Technical report, 1993.

[27] Remote Access (DynDNS). `http://dyn.com/remote-access/`.

[28] D. Research. A bakers dozen, 2016 edition. `http://dyn.com/blog/a-bakers-dozen-2016-edition/`.

[29] RIPE NCC. RIPE Atlas. `https://atlas.ripe.net/`.

[30] RIPE NCC Staff. RIPE Atlas: A global internet measurement network. *Internet Protocol Journal*, 18(3), 2015.

[31] Root Server Operators. Events of 2015-11-30, Nov.2015. `http://www.root-servers.org/news/events-of-20151130.txt`.

[32] S. Sarat, V. Pappas, and A. Terzis. On the use of anycast in DNS. In *Computer Communications and Networks, 2006. ICCCN 2006. Proceedings. 15th International Conference on*, pages 71–78. IEEE, 2006.

[33] ThousandEyes. DDoS attack has varying impacts on DNS Root servers. `https://blog.thousandeyes.com/ddos-attack-varying-impacts-dns-root-servers/`.

[34] M. Weinberg and D. Wessels. Review and analysis of anonmalous traffic to A-root and J-root (Nov/Dec 2015). In 24th DNS-OARC Workshop (Apr. 2016). (presentation).